INRIA

CENTRE DE ROCQUENCOURT

Rapports Techniques

N° 73

LE SYSTÈME INTÉGRÉ DE GESTION ET D'ANALYSE DE DONNÉES PEPIN-SICLA

APPLICATION À UNE ÉTUDE ÉPIDÉMIOLOGIQUE

Geneviève JOMIER
Omar KÉZOUIT
Henri RALAMBONDRAINY
Françoise FACY

Août 1986

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Le système intégré de gestion et d'analyse de données PEPIN-SICLA. Application à une étude épidémiologique.

Geneviève Jomier ISEM Université Paris-Sud.
Omar Kézouit ISEM - INRIA Rocquencourt.
Henri Ralambondrainy INRIA Rocquencourt.
Françoise Facy INSERM Le Vésinet.

Résumé *: Dans cet article, nous allons, à partir d'un exemple, montrer l'intérêt de disposer d'outils puissants à la fois pour la gestion et la manipulation des données, et pour l'investigation statistique de ces données. Un outil intégrant un système de gestion de Base de données relationnelles (PEPIN) et une bibliothèque logicielle de méthodes statistiques (SICLA) est présenté à ce propos. Les spécifications auxquelles devraient répondre ce type d'outils sont également précisées.*

Abstract *: The interest of systems which are powerful both for data management and data analysis is shown in the present paper. A prototype of such system is therefore described, while an epidemiological study is performed. Specifications of such systems are also pointed out.*

INSERM : 44 Chemin de Ronde 78110 Le Vésinet. France.
ISEM : Université Paris-Sud. Bat 490. 91405 Orsay Cedex. France.
INRIA : BP 105. 78153 Le Chesnay Cedex. France.

I Introduction.

Dans cet article nous allons montrer, à l'occasion d'une étude épidémiologique, l'intérêt de disposer d'un logiciel intégrant les fonctionnalités d'un système de gestion de base de données relationnelles (stockage des données d'origine et des résultats d'analyse, manipulation des données, recodage, cohérence, etc) et celles d'un logiciel d'analyse des données (description élémentaire statistique, analyse multidimensionnelle).

La pratique de l'analyse de données fait apparaître que, lors d'une étude statistique, une part importante du temps du statisticien est occupée par la préparation des données à analyser. Il s'agit d'abord de saisir et de stocker les données d'origine en effectuant de multiples contrôles de validité afin de limiter au maximum les erreurs. Ces données, rassemblées dans un ou plusieurs tableaux, et traditionnellement stockées dans un ou plusieurs fichiers, vont donner lieu au cours de l'étude statistique à la création de nouveaux tableaux de

fichiers, vont donner lieu au cours de l'étude statistique à la création de nouveaux tableaux de données composés par extraction et recodage. Ce travail est habituellement effectué à l'aide de programmes créés par le statisticien et opérant sur ces fichiers.

Le but du système que nous implantons est de dégager le statisticien de ces tâches longues et fastidieuses en mettant à sa disposition un outil puissant et sûr.

Plusieurs études ont été menées dans ce domaine des Bases de Données Statistiques (Sho 82, Gho 84, McC 82). Un certain nombre de travaux visant à intégrer base de données et statistique ont été menés jusqu'alors. Ils adoptent différentes démarches.

La première approche, la plus ancienne, consiste à développer les capacités de gestion des données des logiciels statistiques (SPSS, GENSTAT, SAS, etc). En effet, pour résoudre des problèmes de gestion de données, les concepteurs de logiciels statistiques ont implanté, au rythme de leurs besoins, des fonctionnalités nouvelles sur les systèmes de gestion de fichiers dont ils disposaient. Cependant ces solutions entrainent la duplication des données, elles contraignent en général l'utilisateur à se préoccuper de l'organisation physique de ses données et à développer des programmes spécifiques de gestion de données.

La seconde approche, symétrique de la précédente, et adoptée par (KI 81), part d'un SGBD conventionnel et en développe les capacités de traitement statistique. Les fonctions statistiques élémentaires (moyennes, variances, fréquences,...) sont exprimées par des opérations ou des séquences d'opérations sur des tables dans un formalisme proche de celui du modèle relationnel de données. Cependant, en dehors de l'analyse de variance, des traitements statistiques plus complexes du type de la recherche de composantes principales, par exemple, ne semblent pas envisagés.

La troisième approche, que nous avons adoptée, part du constat suivant : pour les deux aspects du problème dont il est question ici, l'analyse statistique et la gestion des données, des logiciels spécialisés existent. Il est alors intéressant de s'orienter vers des systèmes intégrant ces deux types de fonctionnalités. On cumule ainsi dans un même logiciel des avantages propres à chacun. Cette approche a été abordée par le système SIR/DBMS (SIR84), construit autour d'un SGBD relationnel et de modules de statistique descriptive à l'intention des cadres d'entreprises.

Pour mettre en évidence l'intérêt pratique du prototype que nous avons implanté,

après avoir donné la méthodologie générale utilisée en Analyse de Données, nous présentons un exemple d'étude réalisée avec ce système. Puis dans une troisième partie, le problème de gestion des données à analyser est abordé et un rappel des caractéristiques principales des Systèmes de Gestion de Bases de Données (SGBD) relationnels et celles de PEPIN, le SGBD relationnel utilisé permettra de montrer leur intérêt dans ce type d'application. Dans la partie suivante, le logiciel d'analyse statistique SICLA, également utilisé dans ce système est présenté. Dans la dernière partie, nous donnerons une présentation détaillée de l'interface entre PEPIN et SICLA. Une annexe en montre l'usage pour une étude réalisée sur les données d'une enquête épidémiologique.

II Méthodologie en Analyse des Données.

On distingue en général pour la conduite d'une étude en Analyse des Données, les étapes suivantes (Fig 1) :

1) Observation et mesure du phénomène.

Dans la phase initiale, après la définition du problème, des mesures sont faites et les données recueillies puis stockées sur supports informatiques en vue de traitement. Des procédures de contrôle, de nettoyage, de codage sont en général utilisées pour la mise en forme des données.

2) Extraction de tableau pour analyse.

A partir des données observées, un sous-ensemble est extrait pour l'analyse. Le tableau rectangulaire à analyser est construit et ses caractéristiques précisées : type du tableau (individus*variables ou distances par exemple), type des variables (mesure ou binaire par exemple), etc.

3) Gestion du tableau.

Après la construction du tableau divers travaux de mise en forme préalable à l'analyse peuvent encore être effectués : génération de nouvelles variables, transformation des données relativement à une métrique, etc.

4) Analyse du tableau.

L'Analyse des Données propose un certain nombre de méthodes permettant la description effective d'un tableau de données. On s'intéressera principalement ici aux méthodes factorielles et aux méthodes de classification.

Dans les méthodes de type factoriel (Ben 73), l'ensemble des individus et des variables est un nuage de points dans un espace de dimension n, où n est grand. Ces méthodes comme l'Analyse en Composantes Principales ou l'Analyse des Correspondances, recherchent des axes, appelés axes factoriels, bien adaptés aux nuages, et donnent des représentations graphiques dans les plans formés par de tels axes, les plans factoriels, conduisant à une perte d'information minimale. L'examen de tels graphiques permet de détecter des groupes d'individus homogènes et des liaisons entre variables.

Dans les méthodes de type classification automatique (Did 79), on distingue les méthodes de partitionnement et les méthodes hiérarchiques.

Les méthodes de partitionnement (Nuées Dynamiques) déterminent automatiquement une partition en classes homogènes de l'ensemble des individus ou des variables. Les éléments d'une même classe sont ceux qui se ressemblent le plus et qui diffèrent des éléments des autres groupes.

Les méthodes hiérarchiques déterminent des partitions emboitées, où le nombre de classes va décroissant avec les niveaux (DLP 82).

Ces différentes méthodes sont utilisées conjointement et de manière complémentaire pour avoir la description la plus complète possible des données.

5) Edition et interprétation des résultats:

Les différents résultats des méthodes (plans factoriels, partitions, hiérarchies) sont dépouillés et, suivant le cas, on peut être amené à modifier le tableau des données initiales, c'est-à-dire revenir à l'étape 3 ou même à l'étape 2 pour procéder à de nouvelles analyses.

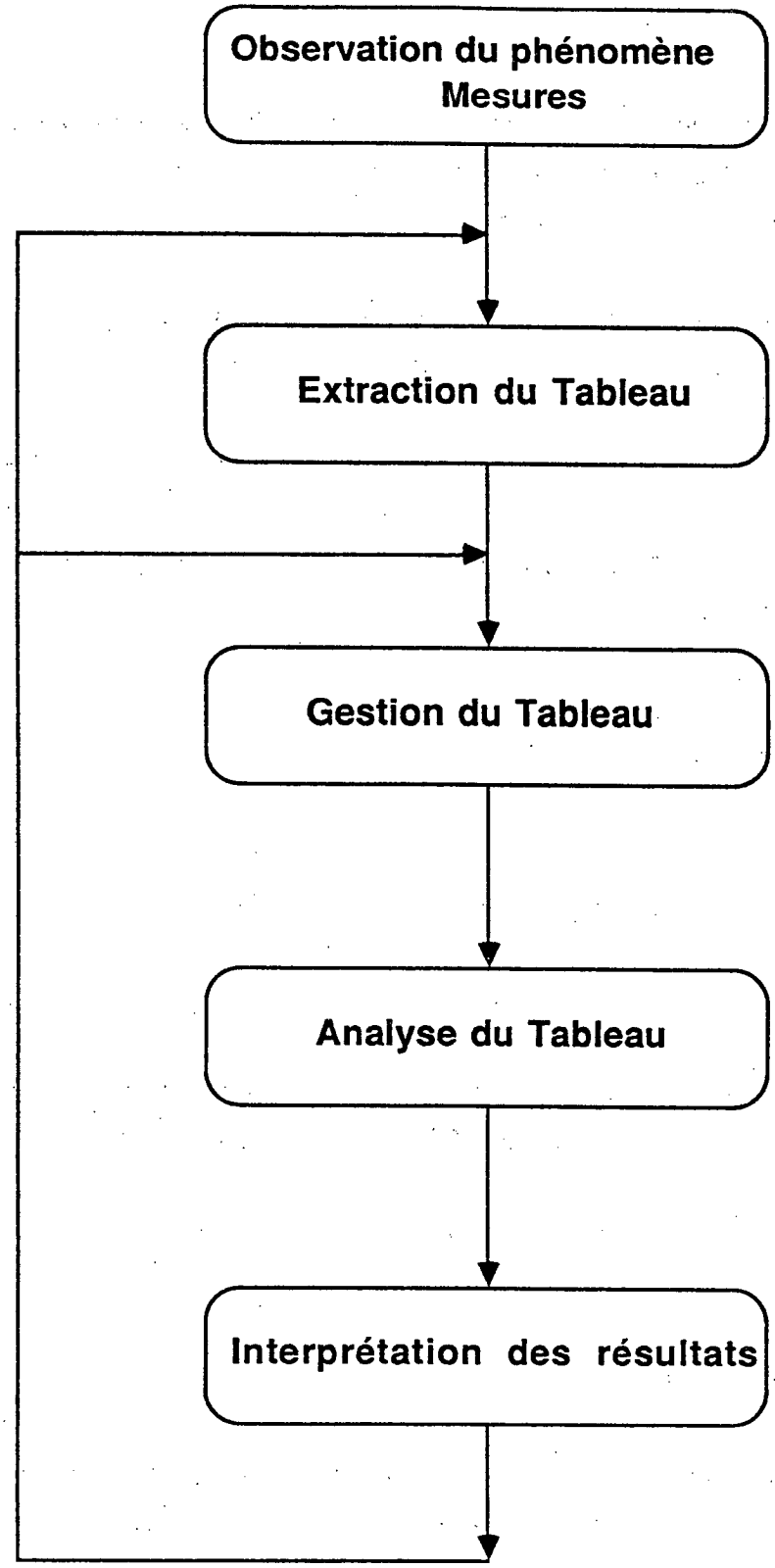

Figure 1 : Méthodologie générale.

III L'ETUDE EPIDEMIOLOGIQUE.

III.1 Présentation des données et de l'objectif de l'étude.

Les données qui vont nous permettre d'illustrer notre propos proviennent d'une étude réalisée en 1978 par l'Institut National de la Santé et de la Recherche Médicale (INSERM) (DFL 84).

Cette enquête porte sur 2088 jeunes de 14 à 20 ans, élèves du second cycle des lycées publics, représentatifs des lycéens à l'échelon national. Elle a été menée dans les régions de Bretagne, des Bouches-du-Rhône et la Région Parisienne.

De nombreux aspects ont été pris en compte au cours de cette enquête. Ils sont organisés en rubriques (groupes de variables). Au total neuf rubriques sont définies à partir d'un ensemble de 163 variables qui décrivent les paramètres socio-démographiques, le cursus scolaire, l'histoire du sujet, le portrait psychologique, les activités, les caractéristiques parentales, la consommation de psychotropes, l'opinion face aux psychotropes, et enfin la santé du sujet.

Les rubriques socio-démographiques et santé, qui vont nous servir au cours de l'illustration, sont présentées à la page suivante, figure 2.

L'objectif de l'étude est la recherche d'une typologie des lycéens selon les variables de santé puis l'interprétation des différentes classes obtenues par rapport à l'arrière-plan socio-démographique des lycéens.

Pour définir des groupes particuliers dans cette population ainsi que des contextes spécifiques impliquant telle ou telle conséquence, la prise en compte simultanée de plusieurs facteurs est nécessaire. Pour cela les méthodes statistiques multidimensionnelles sont tout à fait adéquates (Did 79).

Nous étudions ensuite la stabilité des classes obtenues, en effectuant le même type d'analyse sur une région test, les Bouches du Rhône.

Questionnaire sur l'arrière plan socio-démographique du sujet
rubrique SOCIO-DEMOGRAPHIQUE.

.Quel est l'âge du sujet ?
.Quel est le sexe du sujet ?
.Quelle est la catégorie socio-professionnelle du père ?
.Quelle est la catégorie socio-professionnelle de la mère ?
.Quelle est l'activité du père ?
.Quelle est l'activité de la mère ?
.Quelle est la nationalité du père ?
.Quelle est la nationalité de la mère ?
.Quelle est la situation matrimoniale des parents ?
.Nombre de frères et soeurs du sujet ?
.Dans quelle région vit le sujet ?
.etc

Questionnaire médical : rubrique SANTE.

.Le sujet se considère-t-il comme:
1. bien portant
2. pas très bien portant
3. pas bien portant du tout

.A-t-il eu des accidents durant les 5 dernières années ?
.Le sujet a-t-il été hospitalisé durant les 5 dernières années ?
.Le sujet a-t-il des handicaps importants ?
.Le sujet a-t-il des problèmes psychologiques ?
.Le sujet a-t-il consommé depuis moins d'un an des médicaments :
1. contre la douleur
2. contre l'insomnie
3. contre la fatigue intellectuelle
4. pour maigrir

.etc

Figure 2 : Extraits du questionnaire de l'étude épidémiologique.

III.2 Les différentes étapes de l'étude.

Conformément à la méthodologie présentée en II, la première étape consiste à saisir et à charger les données dans une base gérée par PEPIN, le système de gestion de bases de données (SGBD) utilisé. Ce système est relationnel, c'est-à-dire qu'il opère sur des objets appelés relations, qui sont en fait des tableaux ayant un nombre déterminé de colonnes, appelées attributs, et un nombre quelconque de lignes appelées nuplets.

Chaque rubrique va être saisie comme une relation décrite de la manière suivante :

NOM_DE_RUBRIQUE (NoIndividu, Var1,..., Varn)

où NOM_DE_RUBRIQUE est le nom de la relation, NoIndividu, l'identificateur (ou numéro) de l'individu, et Var1, ..., Varn les n variables composant la rubrique. La relation ainsi définie a par conséquent n + 1 attributs. Ces données sont codées dans l'exemple montré à la figure 3.

Dans la Base de Données relationnelle l'identificateur d'individu est un attribut au même titre que les variables. Il constitue une clé de la relation, c'est-à-dire que chaque valeur de l'identificateur n'apparait qu'une seule fois dans un tableau de données.

SOCIODEMO

Individu	Age	Sexe	.....	Région
1	16	1	.....	1
2	17	2		2
3	16	2		3
4	17	2		2
5	18	1		1
6	18	1		3

SANTE

Individu	Santé	Médicaments	
1	1	4	
2	2	3	
3	1	2	
4	3	4	
5	1	1	
6	1	4	

Figure 3 : Extraits des rubriques de l'enquête épidémiologique.

Lorsque la description des relations de la base de données, qui correspondent aux différentes rubriques, a été donnée au SGBD, les données de ces rubriques peuvent être insérées et contrôlées par le SGBD. La relation SANTE est ensuite transférée dans une structure de données adaptée aux méthodes statistiques de SICLA, le système d'analyse de données utilisé.

Une Analyse des Correspondances Multiples est effectuée sur cette rubrique. La description synthétique des données qu'elle fournit, suggère le nombre de classes et les points initiaux pour une classification automatique. Une partition à six classes est ensuite obtenue en exécutant la méthode des Nuées Dynamiques (Did 79). Cette partition est mémorisée dans la base de données sous forme d'une nouvelle relation dont le schéma est le suivant :

ND_SANTE (NoIndividu, ..., NuméroClasse)

où ND_SANTE est le nom de la relation (Nuées Dynamiques sur la rubrique SANTE pour l'ensemble des individus) et où NuméroClasse est le nouvel attribut indiquant pour chaque individu la classe de santé à laquelle il appartient.

Il faut ensuite faire le lien entre cette partition sur les variables de santé et les variables socio-démographiques. Pour ne conserver que les colonnes qui nous intéressent dans la suite du traitement, on crée une nouvelle relation :

PARTITION_SANTE (NoIndividu, NuméroClasse)

à partir de ND_SANTE. Cette suppression de colonnes d'une relation est une opération relationnelle appelée projection.

Le lien pour un individu entre sa classe de santé et ses variables socio-démographiques est réalisé par une autre opération relationnelle appelée JOINTURE.

La jointure de PARTITION_SANTE (NoIndividu, NuméroClasse) et SOCIODEMO (NoIndividu, Age, ..., Région) sur l'attribut NoIndividu consiste à créer une nouvelle relation PARTITION_SOCIODEMO (NoIndividu, ..., Région, Numéroclasse) dont chaque nuplet est la concaténation d'un numéro d'individu, de ses valeurs pour les variables socio-démographiques et de sa classe de santé. Un exemple de cette opération est donné à la figure 4.

Remarque : Cette opération de concaténation de tableaux est réalisée par une commande du SGBD relationnel alors qu'elle aurait nécessité l'écriture de programmes spécifiques dans la plupart des logiciels statistiques actuels.

La seconde étape de l'enquête épidémiologique va réaliser une analyse similaire sur les seuls lycéens et lycéennes des Bouches-du-Rhône (région marseillaise), afin d'apprécier la stabilité des résultats.

La séquence des opération à réaliser est analogue à celle présentée précédemment pour l'ensemble des lycéens, mais on sélectionne ici les seuls jeunes marseillais. On effectue donc successivement la sélection depuis la relation SOCIODEMO des marseillais dans une nouvelle relation MARSEILLE ; la projection de MARSEILLE dans une nouvelle relation MARSEILLAIS pour ne conserver que l'attribut Noindividu (le seul qui soit utile pour la suite) ; la jointure des relations SANTE et MARSEILLAIS suivant l'attribut Noindividu, ce qui permet de sélectionner les paramètres de santé des marseillais dans une nouvelle relation appelée SANTE-MARSEILLE ; le transfert de cette relation à SICLA pour analyse statistique ; la classification par la méthode des Nuées Dynamiques sur ces données. On cherche une partition à six classes, en choisissant comme points initiaux pour la méthode, des individus de cette région se trouvant dans différentes classes de la partition précédente (réalisée sur l'ensemble des régions) ; l'intégration de la meilleure partition obtenue par la classification, dans la base de données : ceci crée une nouvelle relation dans la base ; la projection de la nouvelle relation pour ne conserver que les attributs NoIndividu et NumeroClasse (ce dernier désigne pour chaque marseillais la classe à laquelle il appartient dans la nouvelle classification) ; la jointure de cette relation et de la relation SOCIODEMO pour interpréter la partition obtenue par rapport aux variables socio-démographiques. L'ensemble de cette manipulation est présentée à la figure 4.

Remarque : la dernière jointure est réalisée sur deux relations ne comportant pas exactement les mêmes individus. On obtiendra cependant le résultat recherché, c'est-à-dire les individus des Bouches-du-Rhône avec leur paramètres socio-démographiques, puisque seuls sont retenus dans la relation résultant de la jointure les individus dont les numéros figurent dans les deux relations à joindre.

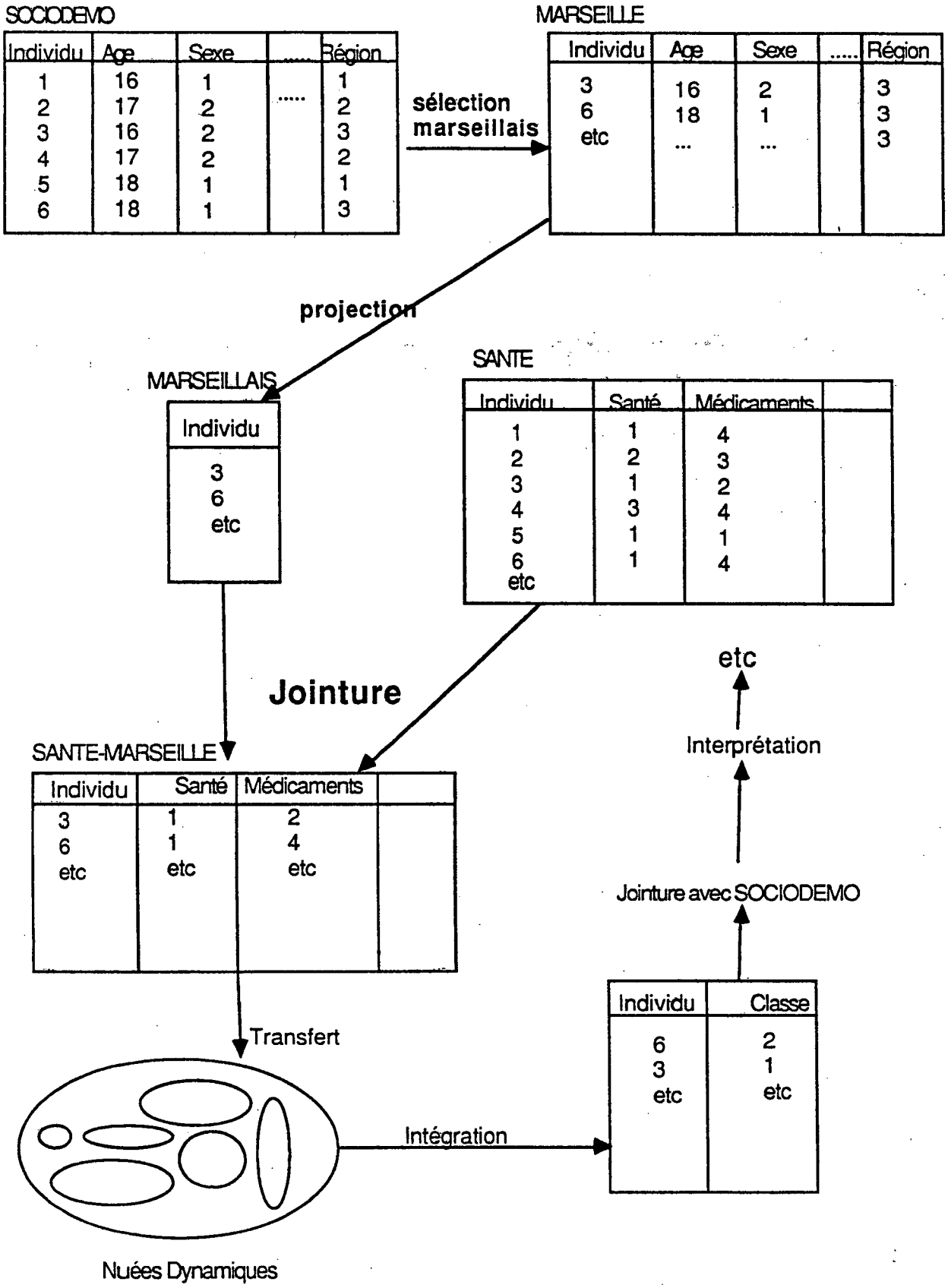

Figure 4 : Etude pour les Bouches-du-Rhône.

III.3 L'interprétation des résultats.

Nous allons maintenant interpréter chacune des partitions et les comparer. SICLA fournit à ce propos des outils d'aide à l'interprétation des partitions, qui allègent la tâche de l'analyste en donnant de façon explicite les profils des classes, ou leurs populations respectives ou encore des indicateurs statistiques divers (pourcentage d'une modalité d'une variable dans une classe, dans la population globale, etc).

La typologie sur l'échantillon total montre 6 classes :

Classe A : représente 65% des sujets.

C'est une classe de lycéens qui se disent bien portants (91%), qui dorment bien et n'ont eu en général ni accident, ni hospitalisation, ni problème grave. Elle est constituée à 47% de jeunes de 15-16 ans, moitié filles, moitié garçons. Leurs parents sont mariés ou vivent ensemble (87%) et les marseillais y sont surreprésentés (33%).

Classe B : représente 19% des sujets.

Les problèmes sont ici diffus. En effet, ni tentation de suicide, ni accident ne caractérise cette classe. En revanche, on trouve des problèmes psycho-somatiques (9%) ou psychologiques (10%), des problèmes graves (10%), des traitements médicamenteux contre l'insomnie (9%) et la nervosité (14%). Cette classe est composée plutôt de garçons (65%), de plus de 18 ans (52%). Le nombre de parents séparés est légèrement supérieur aux autres classes (18%).

Classe C : représente 7% des sujets.

Elle est définie à 100% de jeunes ayant eu des accidents entrainant hospitalisation. Il s'agit plutôt de garçons (66%), plutôt de marseillais (41%). Les plus agés sont plus nombreux que sur l'échantillon total. Les parents sont plus souvent étrangers ici que dans l'ensemble (8% de parents du Maghreb), de même pour les lycéens (13%).

Classe D : représente 4% des sujets.

Elle est constituée à 56% de jeunes ayant tenté de se suicider. Ces jeunes ont des

problèmes psychologiques fréquents (46%), prennent beaucoup de médicaments contre l'insomnie (22%), et la nervosité (39%). Les filles sont surreprésentées (61%), de même que la région parisienne (52%). On note un milieu social dissocié dans 13% des cas.
Classe E (1%) et classe F (4%) :

Ces classes sont caractérisées par des non-réponses à beaucoup de questions médicales et socio-démographiques.

La comparaison de la typologie relative à l'échantillon total à celle relative aux Marseillais montre une stabilité des résultats au niveau de la signification des classes A, C, D, E et F (voir figure 5) avec toutefois des différences au niveau des proportions. Ainsi la population des Bouches-du-Rhône comporte plus de lycéens ayant eu un accident et une hospitalisation : classe C 20% comparé à 7% pour la population totale. Cela s'explique peut-être par une présence plus grande de garcons sportifs dans cette région. La classe D des dépressifs est aussi moins importante dans les Bouches-du-Rhône que dans l'échantillon total. La classe B, quant à elle, ne se retrouve pas ici.

	Typologie 1 Echantillon total	Typologie 2 Echantillon Marseillais
Classe A : bien portants	65%	59%
Classe C : accidents	7%	20%
Classe D : tentative de suicide	4%	1%
Classe E et F : non-réponses	5%	4%

Figure 5 : Correspondance entre l'échantillon total et le sous-échantillon Marseillais

IV La gestion des données à analyser.

L'enquête épidémiologique citée à titre d'exemple au paragraphe précédent a fait apparaître à différents moments de son analyse des besoins de gestion des données que nous allons préciser dans un premier paragraphe. Puis nous verrons de quelle manière les systèmes de gestion de base de données relationnelles peuvent répondre à ces besoins. Enfin nous présenterons une rapide description du SGBD relationnel PEPIN que nous avons utilisé.

problèmes psychologiques fréquents (46%), prennent beaucoup de médicaments contre l'insomnie (22%), et la nervosité (39%). Les filles sont surreprésentées (61%), de même que la région parisienne (52%). On note un milieu social dissocié dans 13% des cas.
Classe E (1%) et classe F (4%) :

Ces classes sont caractérisées par des non-réponses à beaucoup de questions médicales et socio-démographiques.

La comparaison de la typologie relative à l'échantillon total à celle relative aux Marseillais montre une stabilité des résultats au niveau de la signification des classes A, C, D, E et F (voir figure 5) avec toutefois des différences au niveau des proportions. Ainsi la population des Bouches-du-Rhône comporte plus de lycéens ayant eu un accident et une hospitalisation : classe C 20% comparé à 7% pour la population totale. Cela s'explique peut-être par une présence plus grande de garcons sportifs dans cette région. La classe D des dépressifs est aussi moins importante dans les Bouches-du-Rhône que dans l'échantillon total. La classe B, quant à elle, ne se retrouve pas ici.

	Typologie 1 Echantillon total	Typologie 2 Echantillon Marseillais
Classe A : bien portants	65%	59%
Classe C : accidents	7%	20%
Classe D : tentative de suicide	4%	1%
Classe E et F : non-réponses	5%	4%

Figure 5 : Correspondance entre l'échantillon total et le sous-échantillon Marseillais

IV La gestion des données à analyser.

L'enquête épidémiologique citée à titre d'exemple au paragraphe précédent a fait apparaître à différents moments de son analyse des besoins de gestion des données que nous allons préciser dans un premier paragraphe. Puis nous verrons de quelle manière les systèmes de gestion de base de données relationnelles peuvent répondre à ces besoins. Enfin nous présenterons une rapide description du SGBD relationnel PEPIN que nous avons utilisé.

IV.1 Les besoins de gestion des données en Analyse des Données.

Avant de détailler ces besoins, notons que maintenant qu'il disposent de logiciels puissants pour analyser leurs données les statisticiens disent passer 80% de leur temps, lors du traitement d'une enquête, à gérer leurs données. Ces différentes tâches ont été regroupées en quatre catégories.

1) Saisie sur support informatique et contrôle des données d'origine figurant précédemment sur les bordereaux de papiers de l'enquête. Ces données sont généralement codées. Cette procédure génère un ou plusieurs tableaux individus-variables.

2) Extraction de sous-tableaux en vue d'analyse statistique élémentaire ou multidimensionnelle. On peut ne s'intéresser qu'à certaines colonnes, c'est-à-dire vouloir effectuer une analyse seulement sur certaines variables et/ou vouloir éliminer certaines lignes, par exemple des individus dont les questionnaires sont entachés de trop d'erreurs ou de non réponses.

3) Les résultats d'analyse de certains tableaux peuvent entrainer des recodages de certaines variables.

4) Enfin, lorsqu'une analyse a donné des résultats intéressants, on peut vouloir les garder et éventuellement les réutiliser, par exemple pour préparer d'autres sous-tableaux à analyser.

Ces différentes tâches, dont souvent au moins une partie nécessite le développement de programmes ad hoc, sont longues et fastidieuses. Elles nécessitent tests et vérifications. Mais bien souvent, cet effort de programmation ne peut être réutilisé ultérieurement. En effet tous les programmes développés à l'occasion d'une étude sont très liés à la structure physique des données qu'ils manipulent, leur organisation ainsi que la structure de leurs enregistrements et des différents champs des enregistrements (notamment pour les fichiers).

Ces difficultés, communes à tous ceux qui doivent manipuler (stocker, consulter, mettre à jour) des données de volume important, ont amené, à partir de la fin des années 60, au développement de Systèmes de Gestion de Bases de Données. Le but de ces systèmes est précisément d'offrir aux utilisateurs des outils permettant d'effectuer des opérations puissantes sur les données volumineuses sans que l'utilisateur ait à se préoccuper de la structure physique de ses données. Cependant il a fallu attendre la seconde génération des SGBD, les systèmes relationnels, apparue dans les années 80, pour avoir des systèmes capables d'effectuer les opérations présentées au début de ce paragraphe, et bien d'autres encore.

Pour préciser, nous allons indiquer d'abord les caractéristiques des SGBD relationnels puis nous détaillerons certains éléments de PEPIN, le SGBD que nous avons utilisé.

IV.2 Principales caractéristiques des SGBD relationnels.

Comme cela a été signalé au paragraphe III.2, un SGBD relationnel opère sur des relations qui sont en fait des tables ayant un nombre fixé de colonnes, nommées attributs et un nombre quelconque de lignes, les nuplets (Ull 82, Gar 83, Dat 83, PEP 85).

Le créateur d'une base de données définit toutes les relations qu'il souhaite créer, en précisant pour chaque attribut le domaine des valeurs qu'il peut prendre : ainsi l'identificateur d'un individu est une chaine de 4 caractères, l'âge est un entier compris entre 14 et 20, etc. Cette définition est insérée dans le SGBD à l'aide d'un module spécifique dit "de définition de la base". Elle constitue le schéma de la base de données.

A partir du moment où le système connait le schéma d'une base de données il est possible d'y insérer des nuplets, de manière conversationnelle ou par exemple, à partir de fichiers.

Les SGBD relationnels permettent d'effectuer des opérations de consultation et de mise à jour des relations. Les mises à jour de relations sont l'insertion, la suppression et la modification des nuplets d'une relation satisfaisant à une condition pouvant ètre très complexe. Cette dernière opération permet de réaliser les codages et recodages, fréquents en Analyse des Données.

Les opérations relationnelles, utilisées pour consulter la base, ont pour résultat de nouvelles relations qui peuvent être intégrées de manière temporaire ou définitive à la base.

Les opérations relationnelles unaires consistent à extraire d'une relation des colonnes (projection) ou des lignes caractérisées par une condition sur les attributs (sélection).

Les opérations binaires sont des opérations ensemblistes sur les relations : intersection, union, différence, produit cartésien, division. On y ajoute la jointure qui crée une relation à partir de deux autres ayant un attribut de descriptif identique dit attribut de

jointure. Elle consiste à générer les nuplets de la relation résultat en concaténant deux nuplets issus respectivement de chacune des relations opérandes, à condition que l'attribut de jointure ait même valeur dans les deux nuplets (cf paragraphe III.2 et Figure 4 pour une illustration).

IV.3 Le SGBD relationnel PEPIN.

Pour réaliser le prototype d'interface de logiciels d'analyse de données et de gestion de base de données, nous avons choisi d'utiliser la troisième version du SGBD relationnel PEPIN développé sous Unix au Laboratoire ISEM de l'Université Paris-Sud (BEJ 85).

Ce logiciel, actuellement opérationnel et diffusé sur différentes machines, est écrit en Pascal comporte 15000 lignes de source dans sa version de base.

Il comporte principalement le module de définition du schéma et le module de manipulation des données. Ces modules ont été intégrés de sorte que le schéma d'une base de données est lui-même stocké dans la base, dans des relations particulières dites "métarelations".

De ce fait, l'ensemble des relations, c'est-à-dire à la fois les données et leur description, est contenu dans un seul fichier. Ce choix d'implantation est très important pour le maintien de la cohérence des données de la base en cas de panne ou dans un contexte multiutilisateurs. Le fichier contenant la base est un fichier de pages, toutes de même taille. Certaines de ces pages contiennent des nuplets que le SGBD code conformément au schéma de la base. D'autres pages contiennent des informations utilisées pour la gestion des pages de nuplets, les chemins d'accès aux nuplets, les catalogues, etc. C'est cette technique de codage des nuplets par le SGBD conformément au schéma qui permet aux utilisateurs de manipuler les données en ignorant leur structure physique.

Le logiciel lui-même est structuré en couches, chaque couche réalisant une fonction spécifique utilisée par la couche qui lui est immédiatement supérieure. La couche de plus bas niveau gère les échanges de pages entre la mémoire centrale et le fichier qui contient la base. Puis vient une couche qui gère le mécanisme d'adressage, c 'est-à-dire permet de retrouver des pages de nuplets dans le fichier. La couche supérieure gère les nuplets dans les pages de nuplets. La quatrième couche est capable d'exécuter les opérations relationnelles en accédant aux nuplets à l'adresse en mémoire principale indiquée par la couche précédente. La cinquième couche interprète les requêtes posées par les utilisateurs et les code. Enfin la dernière couche gère l'interface avec les utilisateurs. Dans le système intégré PEPIN-SICLA cette interface est

conversationnelle, et procède par menus hiérarchisés.

Une description plus complète des diverses fonctionnalités et du logiciel PEPIN pourra être trouvée dans (PEP 85) ou dans (BEJ 85) et (BEJ 86). Cependant signalons que le système gère automatiquement des transactions. Elles permettent le maintien de la cohérence des données en cas de panne affectant le contenu de la mémoire centrale, dans le contexte multiusagers, et limitent l'incidence d'erreurs de manipulation en offrant toujours la possibilité d'abandonner une transaction erronée avec retour automatique à l'état (cohérent) de la base correspondant au début de la transaction.

Cependant la propriété la plus agréable de PEPIN pour la mise en place de ce travail, est que le système est ouvert et permet des adjonctions aisées grâce en particulier à sa structuration en couches.

V LE LOGICIEL D'ANALYSE DES DONNEES SICLA.

V.1 Les fonctions de SICLA.

Lorsque les données ont été enregistrées, contrôlées, et mises en forme par le SGBD PEPIN, elles sont transmises à SICLA pour l'analyse statistique. Celui-ci est un ensemble de modules permettant la réalisation de l'étude statistique préliminaire, des analyses multidimensionnelles, et l'archivage des résultats d'analyse.

On trouve dans SICLA (Ral 83) des modules d'analyse statistique élémentaire permettant la description des variables quantitatives (calculs de moyennes, histogrammes, corrélations) et des variables qualitatives (tris à plats, tris croisés, calculs de chi-deux). Ainsi pour l'enquête épidémiologique, les tris à plat et croisés ont permis de détecter les questions à fort taux de non-réponse, les liaisons éventuelles entre variables. A partir de telles descriptions, il est possible d'éliminer les questions peu intéressantes, ou de regrouper des modalités à faible effectif.

Alors que les modules d'analyse statistique élémentaire permettent des descriptions univariées (variables par variables) ou bivariées (étude des liaisons entre couples de variables), les méthodes multidimensionnelles permettent de tenir compte simultanément de plusieurs paramètres.

Sicla comporte des outils d'analyse relatifs à la classification (partitionnement et hiérarchies), l'analyse factorielle et la discrimination. Parmi les méthodes de classification

implantées, citons les nuées dynamiques relatives aux variables quantitatives ou qualitatives et la classification sur un tableau de distances. Pour l'analyse factorielle, l'interface avec le logiciel SPAD (LMT 77) permet d'accéder à un ensemble de méthodes dont l'Analyse en Composantes Principales, l'Analyse des Correspondances Simples et Multiples, etc.

Pour l'enquête épidémiologique, les données sont de nature qualitative. Les Nuées Dynamiques sur variables qualitatives et l'Analyse des Correspondances Multiples ont été utilisées. Les variables relatives à la rubrique santé sont considérées comme actives et celles concernant l'arrière plan socio-démographique comme passives (DFL 84).

Les modules associés aux méthodes nécessitent une mise en forme particulière des données. Ainsi les algorithmes de calcul d'histogrammes travaillent sur les données transposées. Les méthodes de discrimination requièrent que les variables à expliquer soient transmises avant les variables explicatives.

SICLA offre une possibilité de mise en forme locale des données. Les modifications faites peuvent être rendues permanentes en transférant les données dans la base.

Les méthodes d'analyse des données génèrent des objets (partitions, tableaux de distances, etc) qui sont identifiés et sauvegardés. Ils sont souvent utilisés par les modules graphiques ou d'interprétation. Certains de ces objets comme les tableaux des coordonnées des individus dans les plans factoriels ou les partitions sont réutilisables par d'autres analyses ou intéressantes à confronter avec d'autres données. L'interface PEPIN-SICLA permet la réinjection de ces structures sous forme de nouvelles relations dans la base de données. En effet elles se présentent sous forme de tableaux et correspondent donc bien à la structure relation n-uplets * attributs.

V.2 La structure du logiciel.

SICLA est constitué d'une bibliothèque d'utilitaires et d'un ensemble de programmes d'application relatifs aux descriptions élémentaires, aux analyses statistiques multidimensionnelles et aux différentes tâches de gestion et d'édition. L'accent est mis sur le contrôle du "bon emploi de l'information". L'architecture est donc régie par un ensemble de règles relatives à divers contrôles d'adéquation entre les types de données et les modules. Ainsi un programme de sélection de variables opérant sur un tableau de données individus*variables ne pourra pas s'exécuter sur un tableau de distances. Ces contrôles permettent la protection des objets du système contre des manipulations intempestives. De même est contrôlée la conformité

entre les types de données statistiques et les analyses. Il n'existe en effet pas de méthode universelle mais des méthodes plus ou moins bien adaptées à des tableaux particuliers. Des vérifications sont donc faites avant l'exécution d'une analyse proprement dite pour éviter l'utilisation abusive de méthodes.

SICLA a été développé dans le projet Classification et Reconnaissance des Formes, de l'INRIA sur le système MULTICS. Il est écrit en FORTRAN 77 et porté sur un certain nombre de machines (VAX, HP9000, UNIVAC, IBM).

En plus de l'interface avec le SGBD relationnel PEPIN, SICLA dispose d'une interface avec un logiciel graphique basé sur les normes GKS permettant l'édition sur console graphique de biplots, courbes, plans factoriels, et d'une interface avec un Système Expert CLAVECIN (DQR 85), permettant une assistance intelligente à l'utilisation de SICLA.

VI PRESENTATION DE L'OUTIL INTEGRE.

VI.1 Le point de vue de l'utilisateur.

PEPIN-SICLA se présente à l'utilisateur comme un logiciel unique (JKR 84). Il est interactif et autodocumenté. Le SGBD joue le rôle de processus maitre, à partir duquel toute commande peut être activée. Il affiche à chaque étape, l'ensemble des possibilités. Le menu principal est divisé en trois groupes de commandes, pour la commodité de la présentation.

Le premier groupe de commandes est constitué des commandes relationnelles. On y trouve principalement les opérations de mise à jour des relations (insertion, suppression, modification de nuplets), les opérations de manipulation des relations (projection, sélection, jointure, tri, union, intersection, différence, destruction) qui donnent lieu préalablement à des contrôles sur les droits d'accès aux relations, des opérations de manipulation du schéma de la base (module de définition des relations et domaines) et enfin les opérations qui permettent de valider ou abandonner une transaction sur la base de données.

Les commandes qui permettent de mettre en oeuvre les méthodes statistiques composent le deuxième groupe. Ces commandes décrites précédemment dans la présentation de SICLA, sont organisées en thèmes, chaque thème correspondant à un ensemble de méthodes. On trouve les thèmes classification, partitions hiérarchiques, discrimination, description, analyse factorielle, etc.

Les commandes particulières de l'interface composent le troisième groupe de

commandes et permettent de gérer les échanges entre PEPIN et SICLA. Elles réalisent le "pont" entre les structures de données statistiques et la base de données relationnelle et représentent le passage obligé entre la manipulation d'objets relationnels et la manipulation d'objets statistiques. Il s'agit particulièrement des commandes d'intégration de données à la base et de transfert de données de la base relationnelle vers le logiciel statistique.

Le prototype actuellement opérationnel gère les tableaux de type individus * variables (tableaux de données brutes et tableaux dérivés). Les matrices (matrice de covariances, de corrélations, de contingence, etc) qu'utilise parfois le statisticien ne sont pas prises en compte par l'interface, bien qu'elles puissent formellement être stockées dans la base. La question qui est posée pour ces objets, est d'évaluer l'intérêt de les ranger dans la base relationnelle, et ce en fonction de l'utilisation ultérieure éventuelle.

Dans la suite de cette partie, les commandes de transfert et d'intégration sont présentées de manière plus détaillée. Auparavant, la base de donnée type est décrite. La figure 6 présente l'architecture générale du système intégré.

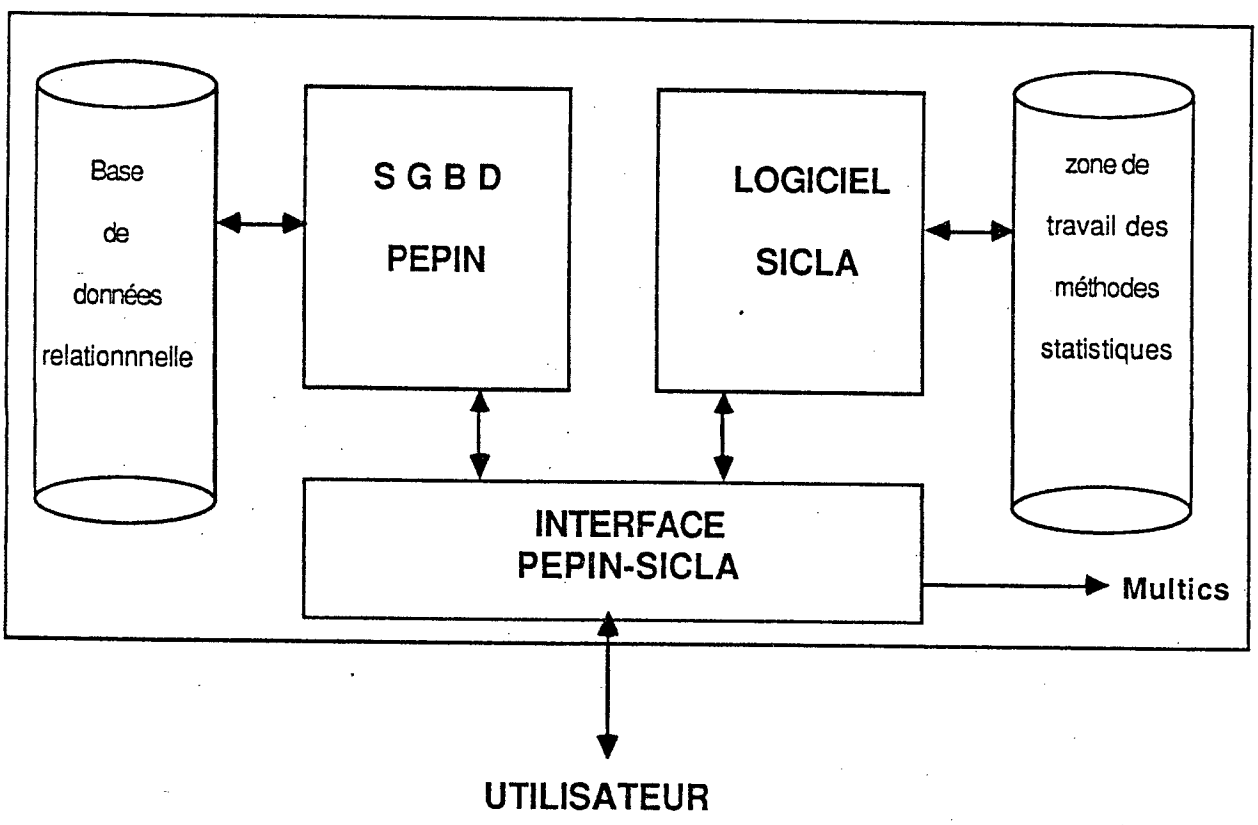

Figure 6 : Architecture générale du système intégré

VI.2 Organisation et chargement de la Base.

VI.2.a Organisation des données dans la base.

Les données sont rangées dans des relations dont le schéma est le suivant :

DONNEE1 (NoIndividu, Var1, Var2, ..., Varn)

ou' l'attribut NoIndividu identifie un individu (ou observation ou sujet) et ou' les attributs Var1 à Varn correspondent aux variables mesurées pour ces individus.

La description des caractéristiques statistiques des variables est contenue dans la base de données, sous la forme de deux relations qui composent le dictionnaire. La première relation, appelée VARIABLES, contient la description des variables, et la seconde, appelée MODALITES, contient une description succincte des modalités des variables qualitatives. Ces deux relations ont le schéma suivant :

DICTIONNAIRE (ID-VARIABLE, LIBELLE, TYPE, NB-MODALITES)

ID-VARIABLE : Identificateur de la variable.
LIBELLE : Libellé de la variable.
TYPE : Type statistique de la variable.
NB-MODALITES : Nombre de modalités de la variable.

MODALITES (ID-VARIABLE, ID-MODALITE, LIBELLE)

ID-VARIABLE : Identificateur de la variable.
ID-MODALITE : Identificateur de la modalité.
LIBELLE : Libellé de la modalité.

Ces deux relations sont créées et maintenues à jour par le système intégré tout au long de la vie de la base.
Trois remarques peuvent être faites ici :

1) La description succincte des modalités contenue dans la relation MODALITES est surtout destinée à l'information de l'utilisateur et à SICLA. L'information la plus importante, et qui permet au SGBD de contrôler la validité des valeurs lues pour chaque variable, est contenue dans la description des domaines (description contenue dans d'autres métarelations de la base). Cette organisation permet de préserver la portabilité de l'interface.

2) Bien que plusieurs relations contenant des données (i-e des tableaux de données) puissent appartenir à une même base, il n'existe qu'un seul dictionnaire contenant la description de

toutes les variables de tous les tableaux de données et toutes leurs modalités. Dans la version actuelle, une variable statistique est reconnue, au niveau du système intégré, par son identificateur, indépendamment de la relation dans laquelle elle apparait. Pour simplifier, on dira qu'une variable est relative à une base de données. De ce fait, il n'y a pas identité entre la notion de variable et la notion d'attribut ; dans le modèle relationnel, un attribut est rattaché à une relation.

Par conséquent, deux variables différentes d'une même base ne peuvent avoir le même identificateur (ID-VARIABLE). L'attribut ID-VARIABLE est donc une clé (au sens du modèle relationnel) de la relation VARIABLES. A titre d'exemple, si CSP désigne la catégorie socio-professionnelle et si elle apparait dans le schéma de plusieurs relations contenant des données statistiques, dans chaque cas CSP représente cette même variable.

3) Les deux relations VARIABLES et MODALITES définissent un objet unique d'un point de vue sémantique, le dictionnaire. Ceci implique un lien de dépendance entre ces deux relations, car les modalités décrites dans MODALITES correspondent à des variables décrites dans VARIABLES. MODALITES fait donc référence à VARIABLES. Ce lien entre les contenus des deux relations, est appelé contrainte d'intégrité référentiel (Dat 81). Il s'exprime ici par l'assertion "toute modalité apparaissant dans le dictionnaire correspond à une variable connue de ce dictionnaire". Des module spécifiques du SGBD gèrent ces contraintes et contrôlent leur validité tout au long de l'utilisation de la base.

VI.2.b Chargement de la base.

Pour le chargement de la base, deux cas fréquents se présentent. Soit les données sont déjà dans une structure de données propre à l'analyse statistique et alors l'opération de chargement de la base se réduit à l'utilisation de la commande d'intégration que nous allons détailler dans les lignes qui suivent (pargraphe VI.3.b). Soit, second cas, les données sont sur d'autres supports (bordereaux de papier ou fichiers informatiques externes) et alors le chargement peut être fait interactivement ou à partir des fichiers contenant les données.

Néanmoins, signalons que pour avoir accès aux méthodes statistiques, il faut que les données soient dans des relations de la base, et que les variables statistiques relatives à ces opérations soient décrites dans le dictionnaire. Ces deux opérations, initialisation du dictionnaire et insertion des données, que nous décrivons maintenant, donnent lieu à des contrôles. L'ordre de leur exécution est indifférent.

1) Initialisation du dictionnaire.

Lors de la création de la base, l'interface crée les deux relations composant le dictionnaire.

L'initialisation du dictionnaire se fait en utilisant une commande particulière de l'interface. Les nuplets correspondant à la description des variables sont insérés dans la relation VARIABLES, puis la description de leurs modalités dans la relation MODALITES. Lorsque l'insertion est terminée, le contrôle de cohérence se déclenche et vérifie que "toute modalité décrite correspond à une variable connue".

Si tel n'est pas le cas, l'utilisateur est averti de l'incohérence et l'opération est annulée. Il doit alors corriger l'erreur (ajouter les nuplets manquants dans la relation MODALITES ou supprimer les conflits d'identificateurs de variables) et recommencer.

A l'issue de cette opération le dictionnaire est nécessairement dans un état cohérent. Ceci signifie unicité de chaque identificateur de variable dans le dictionnaire (i-e son ID-VARIABLE est unique dans VARIABLES) et le fait que les modalités déclarées référent des variables connues. Le contrôle de la cohérence du dictionnaire, est du ressort de l'interface. La correction des incohérences est du ressort de l'utilisateur.

2) Saisie des données.

Les données lues à l'écran ou dans le fichier désigné, sont insérées dans la relation choisie par l'utilisateur. Au fur et à mesure des lectures, la cohérence de la valeur lue avec le domaine de la variable est contrôlée. Si une incohérence est détectée, la valeur est redemandée à l'utilisateur, dans le cas d'une insertion interactive. Si l'insertion se fait à partir de fichier, une incohérence entraine l'arrêt de l'opération et l'envoi d'un message à l'utilisateur.

VI.3 Les échanges entre Base de Données et Structure de Données Statistique.

L'utilisation d'une telle base de données va donner lieu à un va-et-vient continuel entre des requêtes relationnelles c'est-à-dire des requêtes faisant appel aux opérations relationnelles et des requêtes statistiques, c'est-à-dire des requêtes impliquant l'utilisation de méthodes statistiques.

Cependant pour des raisons d'efficacité et d'optimisation, les méthodes statistiques n'opèrent pas directement sur la base de données relationnelle, mais sur leur structure de données statistique propre. En effet, une organisation optimale des données du point de vue de la

structure de données statistique est souvent fort peu judicieuse du point de vue de la base relationnelle et inversement. Par conséquent, pour activer une méthode statistique sur des données stockées dans la base relationnelle, ou pour stocker dans la base relationnelle des résultats d'une méthode statistique contenus à l'origine dans la structure de données statistique il faut que ces deux structures puissent se transmettre leurs données. La suite de ce paragraphe détaille ces échanges.

Pour ces opérations un sous-dictionnaire qui aura la même forme que le dictionnaire global est utilisé. Il est composé de deux relations temporaires, qui disparaissent lors de la première validation demandée par l'utilisateur. L'une, de schéma identique au schéma de VARIABLES, est destinée à contenir l'ensemble des variables à transmettre. L'autre, de schéma identique à MODALITES, contiendra les modalités des variables qualitatives de cet ensemble.

VI.3.a La transmission de la base relationnelle vers la structure de données statistique (transfert).

A partir du nom, donné par l'utilisateur, de la relation contenant des données sur lesquelles il veut faire opérer une méthode statistique, le système extrait du dictionnaire le sous dictionnaire relatif à cette relation. Si tout se passe normalement, les trois relations VARIABLES et MODALITES composant le sous-dictionnaire ainsi que la relation contenant les données sont alors stockées dans une seule et même structure de données adaptée aux méthodes de SICLA.

Si la relation contenant les données à analyser a été créée directement par l'utilisateur et non de manière automatique par le système, il peut se produire qu'un ou plusieurs de ses attributs n'apparaisse pas dans le dictionnaire. En effet la base peut contenir des relations qui ne sont pas destinées à un usage statistique. L'interface permet cette "cohabitation" dans la base tout en empêchant les manipulations erronées. Dans ce cas donc, de transmission est abandonnée et un message est envoyé à l'utilisateur. En tout état de cause, ni le dictionnaire, ni les données ne sont affectées.

VI.3.b La transmission de la structure de données statistique vers la base relationnelle (intégration).

A partir du nom, donné par l'utilisateur, de la structure de données statistique à intégrer dans la base, le système va créer dans la Base de Données les deux relations correspondant au sous-dictionnaire relatif aux données à intégrer, puis la relation devant

contenir les données proprement dites. L'intégration se décompose en deux phases correspondant aux deux niveaux principaux pour le contrôle d'intégrité sur le sous-dictionnaire.
La première phase consiste en la récupération des informations sur les variables depuis la structure de données statistique dans les deux relations composant le sous-dictionnaire, puis le contrôle de l'intégrité de ce sous-dictionnaire. Le système s'assure de la validité de la contrainte d'intégrité référentielle mentionnée plus haut ("toute modalité correspond à une variable connue").

Si tout se passe normalement, à la fin de cette phase on a cohérence du sous-dictionnaire, du point de vue de cette contrainte. La phase suivante peut s'engager. Dans le cas contraire, l'opération s'arrête. L'utilisation du sous-dictionnaire permet d'arrêter l'opération immédiatement et sans dommage pour le dictionnaire global et les données. Cette partie du contrôle n'est réalisée que sur le sous-dictionnaire. En effet, c'est là seulement que pourrait apparaitre une incohérence au regard de la contrainte référentielle.

Seconde phase : La mise à jour du dictionnaire de la Base.
i) les variables nouvelles (n'appartenant pas au dictionnaire global) sont sélectionnées et insérées dans la relation VARIABLES. La deuxième contrainte (unicité de l'identificateur de variable) est alors contrôlée. L'opération n'est réellement effectuée, que si aucun conflit entre les variables n'est signalé. Si au contraire des conflits apparaissent, l'opération est arrêtée sans dommage ni pour le dictionnaire ni pour les données.
ii) les modalités des variables nouvelles sont sélectionnées et insérées dans la relation MODALITES. A l'issue de ces deux phases, la relation devant contenir les données proprement dites est générée et leur insertion avec contrôle sur la cohérence des valeurs lues est effectuée.

En cas de problème au cours de l'opération, un message en clair est envoyé à l'utilisateur et l'opération est annulée. L'utilisateur doit alors corriger l'erreur dans la structure de données statistique et recommencer l'opération.

VII CONCLUSION.

Dans ce papier nous avons montré l'intérêt qu'il y a lors d'un traitement d'enquête à utiliser un système intégrant un logiciel d'analyse de données et un SGBD relationnel.

Le système que nous avons mis en place est opérationnel sur le système MULTICS de l'INRIA. Il donne la possibilité à un utilisateur de manipuler ses données, d'effectuer dessus des

contenir les données proprement dites. L'intégration se décompose en deux phases correspondant aux deux niveaux principaux pour le contrôle d'intégrité sur le sous-dictionnaire.
La première phase consiste en la récupération des informations sur les variables depuis la structure de données statistique dans les deux relations composant le sous-dictionnaire, puis le contrôle de l'intégrité de ce sous-dictionnaire. Le système s'assure de la validité de la contrainte d'intégrité référentielle mentionnée plus haut ("toute modalité correspond à une variable connue").

Si tout se passe normalement, à la fin de cette phase on a cohérence du sous-dictionnaire, du point de vue de cette contrainte. La phase suivante peut s'engager. Dans le cas contraire, l'opération s'arrête. L'utilisation du sous-dictionnaire permet d'arrêter l'opération immédiatement et sans dommage pour le dictionnaire global et les données. Cette partie du contrôle n'est réalisée que sur le sous-dictionnaire. En effet, c'est là seulement que pourrait apparaitre une incohérence au regard de la contrainte référentielle.

Seconde phase : La mise à jour du dictionnaire de la Base.
i) les variables nouvelles (n'appartenant pas au dictionnaire global) sont sélectionnées et insérées dans la relation VARIABLES. La deuxième contrainte (unicité de l'identificateur de variable) est alors contrôlée. L'opération n'est réellement effectuée, que si aucun conflit entre les variables n'est signalé. Si au contraire des conflits apparaissent, l'opération est arrêtée sans dommage ni pour le dictionnaire ni pour les données.
ii) les modalités des variables nouvelles sont sélectionnées et insérées dans la relation MODALITES. A l'issue de ces deux phases, la relation devant contenir les données proprement dites est générée et leur insertion avec contrôle sur la cohérence des valeurs lues est effectuée.

En cas de problème au cours de l'opération, un message en clair est envoyé à l'utilisateur et l'opération est annulée. L'utilisateur doit alors corriger l'erreur dans la structure de données statistique et recommencer l'opération.

VII CONCLUSION.

Dans ce papier nous avons montré l'intérêt qu'il y a lors d'un traitement d'enquête à utiliser un système intégrant un logiciel d'analyse de données et un SGBD relationnel.

Le système que nous avons mis en place est opérationnel sur le système MULTICS de l'INRIA. Il donne la possibilité à un utilisateur de manipuler ses données, d'effectuer dessus des

requêtes du type base de données (ex : Quel est l'ensemble des individus remplissant telle et telle condition ?). Il lui permet en même temps de réaliser sur ces données des études statistiques, le libère des préoccupations physiques d'organisation des données et assure un certain nombre de contrôles de cohérence à sa place.

SICLA et PEPIN sont à l'origine deux logiciels développés indépendamment, écrits l'un en Fortran et l'autre en Pascal. Ils ont déja l'un et l'autre été portés sur différentes machines (Vax, HP9000, etc). Dans un avenir proche le système intégré PEPIN-SICLA devrait être disponible sur ces matériels. Cependant comme on l'a vu en VI l'architecture de l'interface a été concue de manière à permettre à SICLA d'être éventuellement interfacé à d'autres SGBD relationnels.

Ceci met en évidence l'un des avantages du choix d'architecture logicielle que nous avons fait. Chacun des sous-systèmes PEPIN et SICLA peut évoluer indépendamment de l'autre. Ils peuvent être enrichis de nouvelles fonctions sans perturber le système intégré, et à condition que l'interface, dont les contraintes sont les plus légères possibles, aient été respectée.

De nombreuses études ont été consacrées à la spécification des Bases de Données Statistiques (Ber 81, Ber 82), c'est-à-dire aux trés grosses bases de données (ex : bases de données économiques) qui emmagasinent au fil du temps, des données de plus en plus volumineuses et sur lesquelles on veut appliquer certains traitements statistiques. Pour ces applications le problème principal, et toujours ouvert, est celui l'évolution des structures logiques de la base de données au cours du temps. Le but de notre travail est différent. Il s'agit de mettre à la disposition de l'analyste un outil simple d'usage et puissant pour la gestion et l'analyse de ses données.

REFERENCES

(BEJ 85) F. Boufares, Y. El Kabbaj, G. Jomier, H. Ounally :
"La Version SM90 du SGBD relationnel PEPIN."
Journées SM90. Décembre 85. Versailles.

(BEJ 86) F. Boufares, Y. El Kabbaj, G. Jomier, H. Ounally :
"Manuel de présentation de la version 3 du SGBD relationnel PEPIN." Rapport du Laboratoire ISEM. Juin 1986.
Université Paris-Sud. Bat-490. 91405 Orsay Cedex. FRANCE.

(Ber 81) Proceedings of the First LBL Workshop on statistical Database Management.
Décembre 1981. L. Berkeley Labs. California.

(Ber 82) Proceedings of the Second LBL Workshop on statistical Database Management.
Septembre 1982. L. Berkeley Labs. California.

(Dat 81) C. J. Date :
"Referential Integrity."
Proc. of Int. Conf. on VLDB, 1981, Cannes.

(Dat 83) C. J. Date :
"An Introduction To Database Systems."
Addison-Wesley. Reading, Mass. 1983.

(DQR 85) E. Demonchaux, J. Quinqueton, H. Ralambondrainy :
"CLAVECIN : Un sytème expert en Analyse des Données."
Rapport INRIA No 431. Juillet 85, INRIA Rocquencourt.

(DFL 84) F. Davidson, F. Facy, Y. Lechevallier, H. Ralambondrainy :
"Typologie de l'usage de drogues chez les lycéens."
Data Analysis and Informatics. E. Diday Eds.
Elsevier, North Holland. 1984.

(DLP 82) E. Diday, J. Lemaire, J. Pouget, F. Testu :
"Eléments d'Analyse de Données."
Editions Dunod, Paris 1982.

(Gar 83) G. Gardarin :
"Bases de Données : les systèmes et leurs langages."
Ed. Eyrolles PARIS 1983.

(Gho 84) S.P. Ghosh :
"Statistics Metadata : Linear Regression Analysis."
IBM Research Report 4444. San Jose, CA.

(JKR 84) G. Jomier, O. Kezouit, H. Ralambondrainy :
"A System Integrating Data Analysis and Relational Database Management."
Proc. of Int. Conf. on COMPSTAT. August 1984, Vol I, PRAGUE.

(KI 81) Y. Kobayashi, H. Ikeda :

"Additional Facilities of a Conventional DBMS to Support Interactive Statistical Analysis."
Proc. of 1st LBL Workshop on SDBMS. December 1981.

(LMT 77) L. Lebart, A. Morineau, N. Tabard :
"Techniques de la description statistique."
Ed. Dunod PARIS 1977.

(McC 82) J.L. Mc Carthy :
"Metadata Management for Large Statistical Databases."
Proc. of Int. Conf. on VLDB, 1982. Mexico City.

(PEP 85) PEPIN (nom collectif) :
"Introduction aux Systèmes de Gestion de Bases de Données."
Ed. Eyrolles PARIS 1985.

(Ral 83) Henri Ralambondrainy :
"Le système SICLA."
Troisièmes Journées Internationales d'Analyse de Données.
Versailles. Octobre1983.

(SAS) SAS User's Guide, SAS Institute Inc., PO Box 8000
Cary, NC 27511

(SIR 84) "SIR/DBMS reference Manual."
SIR, Inc. 820 Davis St. Suite 400, Evanston, IL 60201.

(Sho 82) A. Shoshani :
"Statistical Databases : Characteristics, Problems and some Solutions."
Proc. of Int. Conf. on VLDB, 1982, Mexico City.

(SPSS) SPSS : Statistical Package for the Social Sciences.
MacGraw Hill, New York.

(Ull 82) J. D. Ullman :
"Principles of Database Systems."
Pitman. 2nd Edition. 1982

ANNEXE : Session avec le systeme integre.

```
Donnez le nom de votre BASE de DONNEES >> ENQUETE

Est-ce une base que vous voulez creer O/N ? n
```

A tous les heureux utilisateurs

Si (par hasard) vous detectiez une erreur en utilisant ce systeme
envoyez un message detaille a Kezouit.Clorec contenant:
1) le nom complet de votre BASE
2) une description precise de l'erreur constatee (A quel moment de la session, etc)
De plus vos suggestions et remarques seront les bienvenues. Envoyez les moi.

Ceci est un systeme integrant un SGBD relationnel (PEPIN) et un logiciel d'analyse statistique (SICLA). Vous pourrez utiliser l'ensemble des methodes statistiques disponibles, si toutefois votre dictionnaire statistique est a jour, c'est-a-dire si les deux relations VARIABLES et MODALITES crees par le systeme, contiennent la description des variables (voir aide pour la commande "initialisation du dictionnaire").

```
Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> ?
```

OPERATIONS SUR LES NUPLETS

insertion	modification	suppression

OPERATIONS SUR LES RELATIONS

affichage	destruction	recherche
classement	union	projection
calcul	jointure	intersection
	difference	

ANALYSE DE DONNEES

transfert_ad	reintegration_ad
creation_relation_donnees	initialisation_dico

toute COMMANDE_SICLA.

OPERATIONS SUR LA BASE

manipulation_du_schema	validation
listage_schema	avortement
aide	arret_session

tapez AU MOINS les quatre premieres lettres, merci.

```
OK >> lister (le schema de la base)

STRUCTURE DE LA BASE sous le format :
  <CREATEUR> RELATION (att1, ..., attn) droits

<SYSTEME> VARIABLES (IDVAR, LIBVAR, TYPESTAT, NBMOD) vli
<SYSTEME> MODALITES (IDVAR, IDMOD, LIBMOD) vli
<KEZOUIT> SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP) vlismd
<KEZOUIT> SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER, NAPR,
 NAMR, MATR, GPP, GPM, FRER, REG) vlismd
```

```
Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> transfert (des donnees pour analyse)

nom de la RELATION contenant les donnees a ANALYSER
sante
nom de l'ATTRIBUT designant les INDIVIDUS dans la RELATION
——> individu

  Fin de transfert

Voulez vous conserver le dictionnaire intermediaire O/N ? o

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> transfert de la seconde relation

nom de la RELATION contenant les donnees a ANALYSER
sociodemo
nom de l'ATTRIBUT designant les INDIVIDUS dans la RELATION
——> individu

 Fin de transfert

Voulez vous conserver le dictionnaire intermediaire O/N ? o

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> lister

STRUCTURE DE LA BASE sous le format :
  <CREATEUR> RELATION (ATT1, ..., ATTn) droits

<SYSTEME> VARIABLES (IDVAR, LIBVAR, TYPESTAT, NBMOD) vli
<SYSTEME> MODALITES (IDVAR, IDMOD, LIBMOD) vli
<KEZOUIT> SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP) vlismd
<KEZOUIT> SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER, NAPR,
 NAMR, MATR, GPP, GPM, FRER, REG) vlismd

Relations temporaires

<KEZOUIT> DICOSANTE (IDVAR, LIBVAR, TYPESTAT, NBMOD) vlismd
<KEZOUIT> MODASANTE (IDVAR, IDMOD, LIBMOD) vlismd
<KEZOUIT> DICOSOCIODEMO (IDVAR, LIBVAR, TYPESTAT, NBMOD) vlismd
<KEZOUIT> MODASOCIODEMO (IDVAR, IDMOD, LIBMOD) vlismd

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> MNDQAL

 ********************************************************************
 * SICLA              Commande MNDQAL              Version 05/05/86  *
 ********************************************************************

 entrez le nom de la structure de donnees choisie

 SANTE

 ————> fichier : SANTE.sdo
```

```
 ensemble des variables qualitatives actives ?

reponse : 1 a 20 (toutes)

 ensemble d'individus actifs ?

reponse : 1 a $ (tous)

 variables selectionnees :

  SANT SOMM MEDC DOUL SOMN NERV FATI MAIG ACCI HOSP
  RAIS HAND NAIS PSYC SCOL PSOM TS   IS   PROB CORP

 L'ensemble des donnees a ete selectionne.


 choix de metrique ?

  1 : metrique du chi-deux relative aux donnees
  2 : metrique euclidienne usuelle

reponse : 1

 choix de l'initialisation pour la methode :

 1 : choix au hasard des points de depart
 2 : les points de depart sont choisis par l'utilisateur
 3 : la partition de depart est choisie au hasard
 4 : la partition de depart est lue sur l'archive

reponse : 1

 nombre de classes pour la partition ?

reponse :  6

 combien d'essais desirez-vous effectuer ?

reponse : 4


 === Fin de MNDQAL      ===

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> INPAQL (interpretation de la partition)

 ************************************************************************
 * SICLA              Commande INPAQL              Version 20/12/85     *
 ************************************************************************

 ensemble des variables qualitatives retenues ?

reponse : 1 a $ (i-e toutes)

 ensemble d'individus retenus ?

reponse : 1a$ (i-e tous)
```

```
 variables selectionnees :

  SANT SOMM MEDC DOUL SOMN NERV FATI MAIG ACCI HOSP
  RAIS HAND NAIS PSYC SCOL PSOM TS   IS   PROB CORP

 L'ensemble des donnees a ete selectionne.

 === Fin de INPAQL      ===

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> CRQLPA

 ************************************************************************
 * SICLA              Commande CRQLPA              Version 15/06/85     *
 ************************************************************************

        creation de variable qualitative
        a partir d'une partition

entrez l'identificateur de la nouvelle variable

reponse :  clas

 === Fin de CRQLPA      ===

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> reintegration (des donnees et de la partition)

nom a donner au TABLEAU de DONNEES a integrer dans la BASE
nd_sante

 votre choix ?

 1. integration d'une structure de donnees
 2. lecture d'un fichier dictionnaire

reponse : 1

Extraction du dictionnaire bien terminee

Voulez vous conserver le dictionnaire intermediaire
O/N ? n

Extraction des donnees lancee...

extraction terminee

Donnees rangees dans la relation ND_SANTE

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> lister

STRUCTURE DE LA BASE sous le format :
  <CREATEUR> RELATION (ATT1, ..., ATTn) droits

<SYSTEME> VARIABLES (IDVAR, LIBVAR, TYPESTAT, NBMOD) vli
<SYSTEME> MODALITES (IDVAR, IDMOD, LIBMOD) vli
<KEZOUIT> SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP) vlismd
```

```
<KEZOUIT> SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER, NAPR,
 NAMR, MATR, GPP, GPM, FRER, REG) vlismd
<KEZOUIT> ND_SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP, CLAS) vlismd

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> projection

nom de la relation :
PROJECTION>>nd_sante
liste des attributs a conserver
PROJECTION>>individu clas
nom de la relation resultat :
PROJECTION>>partition
voulez vous garder le resultat O/N ? o
Quels droits donnez-vous aux autres usagers :
 pressez "?" pour obtenir la liste des droits >> ?
     V)isibilite
     L)ecture
     I)nsertion de nuplets
     S)uppression de nuplets
     M)odification de nuplets
     D)estruction de la relation
taper la liste des droits : vl

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> lister

STRUCTURE DE LA BASE sous le format :
  <CREATEUR> RELATION (ATT1, ..., ATTn) droits

<SYSTEME> VARIABLES (IDVAR, LIBVAR, TYPESTAT, NBMOD) vli
<SYSTEME> MODALITES (IDVAR, IDMOD, LIBMOD) vli
<KEZOUIT> SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP) vlismd
<KEZOUIT> SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER, NAPR,
 NAMR, MATR, GPP, GPM, FRER, REG) vlismd
<KEZOUIT> ND_SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP, CLAS) vlismd
<KEZOUIT> PARTITION (INDIVIDU, CLAS) vlismd

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >>  valider


Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> jointure

nom de la premiere relation :
JOINTURE>>sociodemo
nom de la seconde relation :
JOINTURE>>partition
nom de la relation resultat :
JOINTURE>>part_sociodemo
nom de l'attribut dans la premiere relation :
JOINTURE>>individu
nom de l'attribut dans la seconde relation :
JOINTURE>>individu
```

```
nom de l'attribut dans la relation resultat :
JOINTURE>>individu
Quels droits donnez-vous aux autres usagers :
 pressez "?" pour obtenir la liste des droits >> vl

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> lister

STRUCTURE DE LA BASE sous le format :
  <CREATEUR> RELATION (ATT1, ..., ATTn) droits

<SYSTEME> VARIABLES (IDVAR, LIBVAR, TYPESTAT, NBMOD) vli
<SYSTEME> MODALITES (IDVAR, IDMOD, LIBMOD) vli
<KEZOUIT> SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP) vlismd
<KEZOUIT> SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER, NAPR,
 NAMR, MATR, GPP, GPM, FRER, REG) vlismd
<KEZOUIT> ND_SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
  ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP,  CLAS) vlismd
<KEZOUIT> PARTITION (INDIVIDU, CLAS) vlismd
<KEZOUIT> PART_SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER,
 NAPR, NAMR, MATR, GPP, GPM, FRER, REG, CLAS) vlismd

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> destruction

nom de la relation a detruire ?
DESTRUCTION>>nd_sante
voulez vous vraiment detruire cette relation O/N ? DESTRUCTION>>o
Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> lister

STRUCTURE DE LA BASE sous le format :
  <CREATEUR> RELATION (ATT1, ..., ATTn) droits

<SYSTEME> VARIABLES (IDVAR, LIBVAR, TYPESTAT, NBMOD) vli
<SYSTEME> MODALITES (IDVAR, IDMOD, LIBMOD) vli
<KEZOUIT> SANTE (INDIVIDU, SANT, SOMM, MEDC, DOUL, SOMN, NERV, FATI, MAIG,
 ACCI, HOSP, RAIS, HAND, NAIS, PSYC, SCOL, PSOM, TS, IS, PROB, CORP) vlismd
<KEZOUIT> SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER, NAPR,
 NAMR, MATR, GPP, GPM, FRER, REG) vlismd
<KEZOUIT> PARTITION (INDIVIDU, CLAS) vlismd
<KEZOUIT> PART_SOCIODEMO (INDIVIDU, AGE, NAT, SEXE, PERE, MERE, APER, AMER,
 NAPR, NAMR, MATR, GPP, GPM, FRER, REG, CLAS) vlismd

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> transfert
nom de la RELATION contenant les donnees a ANALYSER
part_sociodemo

           creation d'une structure de donnees

nom de l'ATTRIBUT designant les INDIVIDUS dans la RELATION
---> individu

Voulez vous conserver le dictionnaire intermediaire
O/N ? n
```

```
Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> CRPAQL

                        creation d'une partition
                    a partir d'une variable qualitative

 ************************************************************************
 * SICLA                Commande CRPAQL              Version 15/06/85   *
 ************************************************************************

  entrez le nom de la structure de donnees choisie

PART_SOCIODEMO

  fichier : PART_SOCIODEMO.sdo

  rang de la variable qualitative

reponse : 15

 == Fin de CRPAQL      ==

Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> INPAQL (interpretation de la nouvelle partition)

 ************************************************************************
 * SICLA                Commande INPAQL              Version 20/12/85   *
 ************************************************************************

 ensemble des variables qualitatives retenues ?

reponse : 1 a 14 (toutes sauf la variable a interpreter)

 ensemble d'individus retenus ?

reponse : 1 a $

 variables selectionnees :

  AGE  NAT  SEXE PERE MERE APER AMER NAPR NAMR MATR
  GPP  GPM  FRER REG

 == Fin de INPAQL      ==

STOP fin normale
Quand vous ne savez pas ou plus, pressez <RETURN>
OK >> arret
voulez vous valider les travaux en cours avant de terminer O/N ? o

C'est fini, dommage...
```

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique